500,243

(特許協力条約に基づいて公開された国際出願)

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2003 年7 月10 日 (10.07.2003)

PCT

(10) 国際公開番号
WO 03/056451 A1

(51) 国際特許分類[7]: G06F 17/28

(21) 国際出願番号: PCT/JP02/13185

(22) 国際出願日: 2002 年12 月17 日 (17.12.2002)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2001-395618
2001 年12 月27 日 (27.12.2001) JP

(71) 出願人 (米国を除く全ての指定国について): 独立行政法人通信総合研究所 (COMMUNICATIONS RESEARCH LABORATORY,INDEPENDENT ADMINISTRATIVE INSTITUTION) [JP/JP]; 〒184-0015 東京都 小金井市貫井北町 4-2-1 Tokyo (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 内元 清貴 (UCHIMOTO,Kiyotaka) [JP/JP]; 〒184-0015 東京都 小金井市貫井北町 4-2-1 独立行政法人通信総合研究所内 Tokyo (JP). 井佐原 均 (ISAHARA,Hitoshi) [JP/JP]; 〒184-0015 東京都 小金井市貫井北町 4-2-1 独立行政法人通信総合研究所内 Tokyo (JP).

(74) 代理人: 渡邊 敏 (WATANABE,Satoshi); 〒160-0008 東京都 新宿区三栄町 18-20 渡辺特許法律事務所 Tokyo (JP).

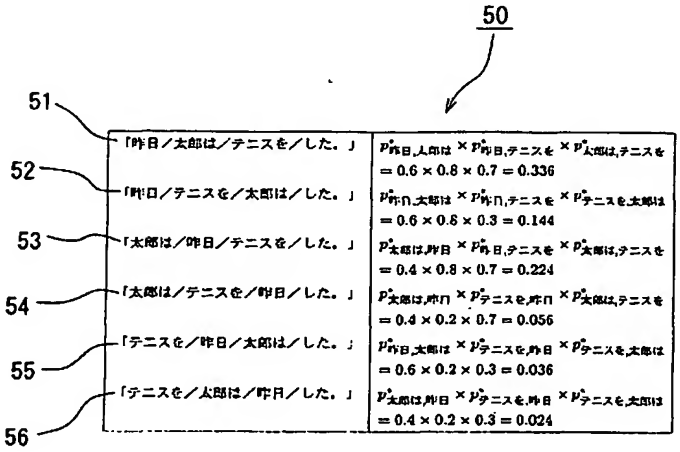
(81) 指定国 (国内): CA, CN, KR, SG, US.

(84) 指定国 (広域): ヨーロッパ特許 (FR, GB).

添付公開書類:
— 国際調査報告書

(54) Title: TEXT GENERATING METHOD AND TEXT GENERATOR

(54) 発明の名称: テキスト生成方法及びテキスト生成装置



51..."YESTERDAY/TARO/TENNIS/PLAYED" P* YESTERDAY,TARO X P* YESTERDAY,TENNIS X P* TENNIS,PLAYED = 0.6 X 0.8 X 0.7 = 0.336
52..."YESTERDAY/TENNIS/TARO/PLAYED" P* YESTERDAY,TARO X P* YESTERDAY,TENNIS X P* TENNIS, TARO = 0.6 X 0.8 X 0.3 = 0.144
53..."TARO/YESTERDAY/TENNIS/PLAYED" P* TARO,YESTERDAY X P* YESTERDAY,TENNIS X P* TARO,TENNIS = 0.4 X 0.8 X 0.7 = 0.224
54..."TARO/ TENNIS/YESTERDAY/PLAYED" P* TARO,YESTERDAY X P* TENNIS,YESTERDAY X P* YESTERDAY,PLAYED = 0.4 X 0.2 X 0.7 = 0.056
55..."TENNIS/YESTERDAY/TARO/PLAYED" P* YESTERDAY,TARO X P* TENNIS,YESTERDAY X P* TARO,PLAYED = 0.6 X 0.2 X 0.3 = 0.036
56..."TENNIS/TARO/YESTERDAY/PLAYED" P* TARO,YESTERDAY X P* TENNIS,YESTERDAY X P* TENNIS,TARO = 0.4 X 0.2 X 0.3 = 0.024

(57) Abstract: A text generating method and a text generator for preparing a natural text from a keyword based on at least one keyword. The keyword is input from a keyword input unit, and the texts and phrases including at least one keyword are extracted from a database by a text and phrase search and extraction unit. A text generation unit performs the morpheme analysis and the syntax analysis of the extracted texts, etc., and outputs the natural texts by combining the keyword with the texts, etc.

[続葉有]

WO 03/056451 A1

(57) 要約:

本発明は、１つ以上のキーワードを基に、そのキーワードから自然なテキストを生成する生成方法・生成装置を提供する。そのために、キーワードをキーワード入力部から入力し、キーワードを１つでも含むテキストや語句をデータベースからテキスト語句検索抽出部で抽出する。テキスト生成部では抽出されたテキストなどを形態素解析・構文解析し、テキストなどにキーワードを組み合わせることで自然なテキストを出力する。

明細書

テキスト生成方法及びテキスト生成装置

技術分野

本発明は自然言語処理方法及び装置に関する。特に、いくつかのキーワードからテキストを生成する手法に特徴を有する。

従来背景

近年、コンピュータによって言語のテキストを解析する技術、或いは生成する技術の開発が進んでいる。特にテキストの生成においては、いかに自然なテキストを生成できるかが、課題となっており、人間が生成したものと遜色のない生成方法の提供が求められている。

例えば、いくつかのキーワードを入力したときに、それらのキーワードを用いて自然なテキストを生成する技術は、外国人など、文章作成の苦手な者への作成支援を行うことに寄与する。

また、単語を列挙することで相手への意思を伝達できるため、機械翻訳に近い使い方も可能である。

例えば、失語症患者の文生成支援では、現在、日本全国でおよそ１０万人程度の失語症患者がおり、その８割程度の人はとぎれとぎれの文（単語の列）を発声できる、あるいは単語の候補を提示してあげると言いたいことを表現するためにその中からいくつか単語を選択することができると言われている。

そこで、例えば「彼女　公園　行った」などを発声あるいは選択し、そこから自然な文「彼女が公園へ行った」、「彼女と公園へ行った」などを生成して提示することによって、患者のコミュニケーションを支援する。

このように、１つ以上のキーワードを入力して、自然なテキストを生成する従来の技術として、テンプレートをもとに文を生成する技術や、キーワー

ドをもとにデータベースから文を検索する技術はすでに存在する。

しかし、これらの技術ではテンプレートに合致する場合のみ、あるいはデータベース中に含まれる文と合致する場合のみにしか有効でなく、いずれも限られた型の文しか生成できない。

また、検索の際、適合しやすくなるようにキーワードを類義語などに置き換える技術も提案されているが、キーワードから生成されるべき文のバリエーションは多岐に亙るため、十分であるとはいえない。

発明の開示

本発明は、このような従来の背景から創出されたものであり、１つ以上のキーワードを基に、そのキーワードから自然なテキストを生成する生成方法・生成装置を提供する。

すなわち、本発明によると、次の各ステップに基づいてテキストの生成を行う。

まず、１個以上のキーワードとなる単語を入力する入力ステップで、「彼女」「公園」「行った」などを入力する。

そして、該キーワードに係るテキスト又は語句を、データベースから抽出する抽出ステップに進む。データベースには多くの例文が搭載されており、例えば「彼女」という単語を含むテキストや語句を探して抽出する。

次に、抽出されたテキスト又は語句を組み合わせて、入力したキーワードを用いる最適なテキストを生成する。このテキスト生成ステップでは、例えば「彼女」、「～へ」、「行った」を含むテキストがデータベース中にあるときに、「彼女は公園へ行った」というように組み合わせてテキストを生成する。

ここで、前記抽出ステップでテキストだけを抽出する構成とし、前記テキスト生成ステップにおいて、抽出されたテキストを形態素解析及び構文解析し、該テキストの係り受け構造を得てもよい。そして、キーワードを含む係り受け構造を形成することによって、より自然なテキスト生成を実現するこ

ともできる。

さらに、キーワードを含む係り受け構造を形成する過程で、係り受けモデルを用いてテキスト全体の係り受け確率を求め、該確率の最大のものを最適なテキストとして生成してもよい。

本発明では、語順についても語順モデルを用いて、自然な文の並びとなるテキストの生成を図ることもできる。テキスト生成ステップにおいて、係り受け構造を形成する過程あるいは形成した後で用いることができる。

また、テキスト生成ステップにおいて、キーワードの全ての配列について、任意の２つのキーワード間に補完すべき単語があるか否かを学習モデルを用いて判定することもできる。学習モデルにおいて補完すべき確率の高い単語から順に補完するとき、いずれのキーワード間についても補完すべき単語がない確率が最も高くなるまで繰り返す。キーワードには補完した単語を編入することができるので、補完された単語間にもさらに補完することもできる。これにより、好適な補完が実現できるので、与えるキーワードが少ない場合でも、自然なテキスト生成を図ることができる。

また、本発明では、上記のデータベースに、特徴的なテキストパターンを有するテキストを備え、テキスト生成ステップがその特徴を反映したテキストを生成する構成をとることもできる。

例えば、文体や言い回しなどについて特徴のあるテキストをデータベースに備えておくことで、生成されるテキストが、その特徴に準拠したテキストになる。

本発明は、文又は文章のテキストを生成するテキスト生成装置として提供することもできる。該テキスト生成装置には、１個以上のキーワードとなる単語を入力する入力手段、複数のテキストで構成されるテキストデータベース、該キーワードに係るテキスト又は語句を、該テキストデータベースから検索し、抽出する抽出手段、抽出されたテキスト又は語句を組み合わせて、入力したキーワードを用いる最適なテキストを生成するテキスト生成手段を備える。

抽出手段がテキストだけを抽出する構成では、抽出されたテキストを形態素解析及び構文解析し、該テキストの係り受け構造を得る解析手段と、前記キーワードを含む係り受け構造を形成する係り受け構造形成手段とをテキスト生成手段に含むこともできる。

特に、テキスト生成手段において、係り受け構造形成手段が、係り受けモデルを用いてテキスト全体の係り受け確率を求め、該確率の最大のものを最適なテキストとして生成するとよい。

テキスト生成手段において、係り受け構造を形成する過程あるいは形成した後で、語順モデルを用いて自然な文の並びとなる最適なテキストを生成することもできる。

また、テキスト生成手段において、前記キーワードの全ての配列について、任意の２つのキーワード間に補完すべき単語があるか否かを学習モデルを用いて判定し、学習モデルにおいて補完すべき確率の高い単語から順に補完するとき、いずれのキーワード間についても補完すべき単語がない確率が最も高くなるまで繰り返す単語補完手段を含んでもよい。

テキスト生成装置においても、上記同様、データベースに特徴的なテキストパターンを有するテキストを備え、テキスト生成手段がその特徴を反映したテキストを生成するようにしてもよい。

さらに、パターン選択手段を設けることで、複数のテキストパターンを適宜選択切換することもできる。

図面の簡単な説明

第１図は、本発明によるテキスト生成装置の説明図である。

第２図は、テキスト生成部において解析された係り受け構造の部分グラフである。

第３図は、テキスト生成部において生成された係り受け構造木である。

第４図は、別の例文における係り受け構造木である。

第５図は、係り文節の順序が適切である確率の計算例である。

符号の指示部位は次の通りである。１：テキスト生成装置、２：入力するキーワード、３：出力されたテキスト、１０：キーワード入力部、１１：テキスト語句検索抽出部、１２：テキスト生成部、１２ａ：解析部、１２ｂ：形成部、１２ｃ：評価部、１３：データベース

発明を実施するための好ましい形態

以下、本発明の実施方法を図面に示した実施例に基づいて説明する。なお、本発明の実施形態は以下に限定されず、適宜変更可能である。

図１には本発明におけるテキスト生成装置（１）の説明図を示す。該装置には、キーワード入力部（１０）、テキスト語句検索抽出部（１１）、テキスト生成部（１２）と共に、データベース（１３）を備える。データベース（１３）には予め複数のテキストがテーブルとして備えられており、該テーブルの内容については適宜変更させることもできる。内容を変更することで様々なテキストの生成を実現できるが、この点については後述する。

そして、例えば「彼女」「公園」「行った」の３つのキーワード（２）をキーワード入力部（１０）から入力すると、テキスト語句検索抽出部（１１）がデータベース（１３）からキーワードの少なくとも１つを含むテキストや語句を検索して、それらを抽出する。

さらにテキスト生成部（１２）では抽出されたテキストや語句に基づき、それらを組み合わせることで、自然なテキスト、ここでは「彼女は公園へ行った」（３）を出力する。

各過程をさらに詳述する。まず、キーワード入力部（１０）において入力されたキーワードについて、テキスト語句検索抽出部（１１）でデータベース（１３）からキーワードｎ個を含む文を抽出する。ここで、キーワードは１つでも含めばよい。抽出された文はテキスト生成部（１２）に送られる。

テキスト生成部（１２）は、解析部（１２ａ）と形成部（１２ｂ）、評価部（１２ｃ）から成り、解析部（１２ａ）においてまず抽出した文の形態素解析及び構文解析を行う。

形態素解析には、例えば本件出願人らが特願２００１－１３９５６３号で出願中のＭＥモデルによる形態素の解析方法を用いることができる。

ここで、形態素解析をＭＥモデルに適用するために、形態素としての尤もらしさを確率として表す。

すなわち、文が与えられたとき、その文を形態素解析するという問題は文を構成する各文字列に、２つの識別符号のうち１つ、つまり、形態素であるか否かを示す「１」又は「０」を割り当てる問題に置き換えることができる。

さらに、形態素である場合には文法的属性を付与するために「１」を文法的属性の数だけ分割する。すると、文法的属性の数がｎ個のとき、各文字列に「０」から「ｎ」までのうちいずれかの識別符号を割り当てる問題に置き換えることができる。

したがって、形態素解析にＭＥモデルを用いた手法では、文字列が、形態素であって、かついずれかの文法的属性を持つとしたときの尤もらしさをＭＥモデルにおける確率分布の関数に適用することで求められる。形態素解析においてはこの尤もらしさを表す確率に、規則性を見いだすことで処理を行っている。

用いる素性としては、着目している文字列の字種の情報、その文字列が辞書に登録されているかどうか、１つ前の形態素からの字種の変化、１つ前の形態素の品詞などの情報を用いる。１個の文が与えられたとき、文全体で確率の積が最大になるよう形態素に分割し文法的属性を付与する。最適解の探索には適宜公知のアルゴリズムを用いることができる。

このように、ＭＥモデルを用いた形態素解析方法は、例えば未知語を含んでいても有効な形態素解析ができるなど、優位性の高い方法である。本発明の実施においては、上記方法によることが特に効果的であるが、必ずしも限定されるものではなく、任意の形態素解析方法を用いることができる。

さらに、解析部（１２ａ）における構文解析についてもＭＥモデルを用いた解析手法を導入することができる。構文解析についても、他の任意の手法

に置き換えることができるが、一実施例として以下の手法を示す。前記データベース（１３）はテキスト生成部（１２）からも参照が可能であり、本ＭＥモデルではデータベースに含まれる複数のテキストから学習を行うことができる。

構文解析のうち、係り受け解析についての導入をする。どの文節がどの文節を修飾するかという日本語の係り受け関係には、主に以下の特徴があるとされている。すなわち、

（１）係り受けは前方から後方に向いている。

（２）係り受け関係は交差しない。（以下、これを非交差条件と呼ぶ。）

（３）係り要素は受け要素を1 つだけもつ。

（４）ほとんどの場合、係り先の決定には前方の文脈を必要としない。

本実施例では、これらの特徴に着目し、統計的手法と文末から文頭に向けて解析する方法を組み合わせることにより高い解析精度を得ることを実現した。

まず、文末から順に２つずつ文節を取り上げ、それらが係り受けの関係にあるかどうかを統計的に決定する。その際、文節あるいは文節間にみられる情報を素性として利用するが、どのような素性を利用するかが精度に影響する。

文節は、前の主辞にあたる部分と後ろの助詞や活用形にあたる部分に分けて考え、それぞれの素性とともに文節間の距離や句読点の有無なども素性として考慮する。

さらに括弧の有無や文節間の助詞「は」の有無、係り側の文節と同じ助詞や活用形が文節間にもあるか否か、素性間の組み合わせについても考慮している。

ＭＥモデルによればこういった様々な素性を扱うことができる。

そして、この方法では決定木や最尤推定法などを用いた従来の手法に比べて学習データの大きさが１０分の１程度であるにも関わらず、同程度以上の精度が得られる。この手法は学習に基づくシステムとして、最高水準の精度

を得られる手法である。

さらに、従来は、学習データから得られる情報を基に、２つの文節が係り受け関係にあるか否かを予測するのに有効な素性を学習していたが、新たに前文節が「後文節を越えて先にある文節に係る」「後文節に係る」「後文節との間にある文節に係る」の３つの状態のどれであるかを予測するのに有効な情報を学習する方法によって、より高精度な係り受け解析を可能にしている。

このように、ＭＥモデルを用いた形態素解析方法、構文解析方法を採用することによって、解析部（１２ａ）ではデータベース（１３）から検索抽出されたテキストを正確に解析し、該テキストの係り受け構造を得る。該係り受け構造は部分グラフとして表すことができる。ここで、グラフ構造のノードが文節、アークが係り受けとする。

各キーワードを少なくとも一つ含む部分グラフをすべて抽出し、頻度を調べる。ノードは汎化した情報(人名、組織名などの固有表現や品詞)のみを持つものも考慮する。

データベース（１３）から上記のキーワードに基づいて抽出され、解析した結果のうち、頻度が高かったものが図２のａとｂである。例えばａにおいて、キーワード「彼女は」をノード（親ノード１）（２０）とすると、「＜名詞＞＋へ」がノード（親ノード２）（２１）、「＜動詞＞。」がノード（子ノード）（２２）として係り受け関係（２３）をもつ。

この過程より先はテキスト生成部（１２）のうち形成部（１２ｂ）における処理に移行する。ただし、本実施例では、テキスト生成部（１２）における解析と形成は以下に示すように一体的な処理であり、相互に連係して動作する。

入力するキーワードｎ個は係り受け関係にあると仮定し、入力単語ｎ個を含むような係り受け構造木を生成する。木の生成には上記の部分グラフを組み合わせて用いる。

例えば、上記のキーワード３個を入力して、それらが係り受け関係にある

と過程し、部分グラフを組み合わせる（この場合は当てはめる）と、図３に示すａ及びｂが得られる。

ここで、生成された２つの木（図３ａ・ｂ）のうち、いずれが適当であるかを再び上記の係り受けモデルを用いて選択する。

順序付けの際には、組み合わせた部分グラフ間での一致する割合、頻度、係り受け関係を考慮する。特にｎが３以上の場合、単語ｎ個間の係り受け関係には曖昧性があるが、曖昧性の解消には、係り受けモデルを利用する。係り受けモデルによって求められる確率値が大きなものを優先して順序付けする。

その結果、ａの木における確率値により高い結果が得られ、最適な係り受け関係はａであることが選択される。

日本語においては、語順の制限が比較的緩やかであり、係り受け関係が決定されると自然なテキストに近い結果が得られるが、本発明の対象とする言語は必ずしも日本語に限られず、他の言語で用いることも考えられる。

また、日本語においてもより自然なテキストに寄与するためには最も自然な語順が選択されることが望ましく、本発明では、次のように並べ替えることができる。

まず、優先順位の高い木から、自然な文の並びに置き換えて出力する。その際、依存構造から自然な並びの文を生成するＭＥモデルを用いた語順モデルを利用する。語順モデルの学習についてもデータベース（１３）を参照して行うことができる。

語順が自由であると言われる日本語でも、これまでの言語学的な調査によると、時間を表す副詞の方が主語より前に来やすい、長い修飾句を持つ文節は前に来やすいといった何らかの傾向がある。もしこの傾向をうまく整理することができれば、それは自然な文を生成する際に有効な情報となる。ここで語順とは、係り相互間の語順、つまり同じ文節に係っていく文節の順序関係を意味するものとする。語順を決定する要因にはさまざまなものがあり、例えば、修飾句の長い文節は短い文節より前に来やすい、「それ」などの文

脈指示語を含む文節は前に来やすい、などがあげられる。

本実施例においては、上記のような要素と語順の傾向との関係、すなわち規則性を所定のテキストから学習する手法を考案した。この手法では、語順の決定にはどの要素がどの程度寄与するかだけでなく、どのような要素の組み合わせのときにどのような傾向の語順になるかということも学習に用いるテキストから演繹的に学習することができる。個々の要素の寄与の度合はＭＥモデルを用いて効率良く学習する。係り文節の数によらず２つずつ取り上げてその順序を学習する。

文を生成する際には、この学習したモデルを用いて、係り受け関係にある文節を入力とし、その係り文節の順序を決めることができる。語順の決定は次の手順で行なう。

まず、係り文節について可能性のある並びをすべて考える。次に、それぞれの並びについて、その係り文節の順序が適切である確率を学習したモデルを用いて求める。この確率は、順序が適切であるか否かの「０」または「１」に置き換え、ＭＥモデルにおける確率分布の関数に適用することで求められる。

そして、全体の確率が最大となる並びを解とする。全体の確率は、係り文節を２つずつ取り上げたときその順序が適切である確率を計算し、それらの積として求める。

例えば、「昨日／テニスを／太郎は／した。」という文で最適な語順の決定を説述する。上記と同様に係り受け構造木を作成すると、最も確率値の高い構造木が図４のように得られる。

すなわち、動詞「した。」（４３）に係る文節は「昨日」（４０）、「テニスを」（４１）、「太郎は」（４２）の３つである。この３つの係り文節の順序を決定する。

図５に係り文節の順序が適切である確率の計算例（５０）を示す。

まず、２個の文節ずつ、すなわち「昨日」と「太郎は」、「昨日」と「テニスを」、「太郎は」と「テニスを」の３つの組み合わせを取り上げ、学習

した規則性によりそれぞれこの語順が適切である各確率を求める。

例えば、図において「昨日」「太郎は」の語順になる確率は「p*（昨日，太郎は）」で表され、その確率は０．６とする。同様に、「昨日」「テニスを」は０．８、「太郎は」「テニスを」は０．７とすると、図５における１段目の語順（５１）の確率は各確率を積算し、０．３３６となる。

次に、６つの語順（５１ないし５６）の可能性すべてについて全体の確率を計算し、最も確率の高いもの「昨日／太郎は／テニスを／した。」（５１）が最も適切な語順であるとする。

同様に、前記したテキスト「彼女は／公園へ／行った。」でばさらに少ない組み合わせの確率を計算することで、「彼女は公園へ行った。」が最も自然で最適なテキストであると求められる。

また、該語順モデルについては、汎化したノードが含まれる場合、そのまま提示することによって、人名や地名、日付などが入り易い場所が分かる。

ここで、上記における語順モデルでは係り受け構造を入力としているが、本発明の実施においては係り受け構造の形成過程においても語順モデルを用いることができる。

以上により、テキスト生成部（１２）の形成部（１２ｂ）では、係り受けモデル、語順モデルなどにより最適と考えられる複数のテキストが候補として形成される。本発明ではこれらをそのままテキスト生成装置（１）から出力することもできるが、以下では、さらにテキスト生成部（１２）に評価部（１２ｃ）を配置し、テキストの候補を評価することにより再順序付けする構成を示す。

評価部（１２ｃ）では、入力されたキーワードの順番や、抽出したパターンの頻度、係り受けモデルや語順モデルから計算されるスコアなど様々な情報を総合してテキストの候補の評価を行う。評価部（１２ｃ）においてもデータベース（１３）を参照することができる。

例えば、キーワードの順番が上位のものについてはより重要なキーワードとして、該キーワードの役割が特に重要な候補中のテキストを、より最適な

テキストとして評価したり、前記では係り受けモデルや語順モデルといったモデル毎に確率を求めたが、それらを勘案して、総合的な評価を行うようにする。

本評価部（１２ｃ）の働きによって、自然なテキストとして形成された候補のうちでも、特に最適と考えられるテキストを例えば順位を付けて複数出力することができるようになる。

本発明によるテキスト生成装置（１）は、さらに別の言語処理システムに導入することも可能であって、このように複数の出力を行っても良いし、上記順位が最も高いものを１つ出力してもよい。

また、順位が一定以上に高いもの、あるいは確率やスコアなどで一定の閾値以上のものを出力し、人手によって選択する構成をとってもよい。

上記評価部（１２ｃ）の構成では形成部（１２ｂ）で形成された候補を入力するのみの構成であるが、さらに評価部（１２ｃ）において複数のテキストからなる文章全文について各テキストの候補のいずれを選択するか、全文の流れから評価し、各テキストの候補から１つを決定してもよい。

この時、文章全文中の少数のテキストが前後の文との整合性において不自然な場合には、再び解析部（１２ａ）や形成部（１２ｂ）における処理に差し戻し、全文に亘って自然なテキストが出力できるように別の候補を形成させるようにしてもよい。

テキスト生成部（１２）によって最適な構文、語順で生成されたテキスト「彼女は公園へ行った。」（３）は以上に説述したテキスト生成装置（１）から出力される。ここでは、最も自然と考えられるテキスト（３）を１つ出力した。

このように、本発明では、１つ以上のキーワード（２）を入力することで、データベース（１３）を参照しながらも、従来の技術とは異なる構成で自然なテキストを生成することができる。

さらに、本発明ではキーワードが十分でない場合の、補完方法についても提供する。

すなわち、キーワードｎ個が入力されたとき、その単語間をＭＥモデルを用いて補完する。モデルに対してはｎ個のうちの２個を入力し、該２キーワード間を補完する。

そして、任意の２キーワード間について、補完すべき単語があるか否か、補完できる単語が複数ある場合には、各単語について生起する確率を求めていく。確率の最も高い単語から順に補完し、すべての２キーワード間についてこの処理を繰り返す。

最後にどの２キーワード間についても「補完しない」が最も確率が高くなるとき補完を止める。

このような補完処理によれば、キーワードが十分に与えられていない場合であっても、ＭＥモデルによってある程度までキーワードを補うことができるので、入力されたキーワードだけでは自然なテキストが生成出来ない場合にも、有効なテキストを出力できるようになる。

本補完方法では、さらに上記テキスト生成部においてテキスト生成に用いることも可能である。

例えば、上記例で示したように、「彼女」「公園」「行った。」が与えられたときに、「彼女」と「公園」の間には「は」「が」「と」などが生起し、その中で最も生起する確率の高い「は」を補完する。

同様に「彼女」と「行った。」の間には「は」「が」「と」などが生起し、ここでも最も確率の高い「は」を補完する。「公園」と「行った」の間では「へ」「に」等が生起し、確率の高い「へ」を補完する。

これらの補完を繰り返して、最終的に全文について補完される確率を算出し、各積算することによって、全文について最も確率の高くなる補完組み合わせを採り、テキストを生成する。この場合には、「彼女は公園へ行った。」となり、前述した本発明に係る方法と同様の効果が得られる。

本発明では、このように前述のテキスト生成方法を基本としながら、キーワードの補完、さらにはこの補完方法を用いたテキストの生成を実現する。

以上のような本発明によるテキストの生成技術は次のような場合に特に好

適に用いることができる。

まず、失語症患者の文生成支援に用いることができる。とぎれとぎれの文(単語の列)、例えば「彼女 公園 行った」、から自然な文を生成し、文の候補、「彼女が公園へ行った」、「彼女と公園へ行った」などを提示する。患者は提示されたテキストを承認するだけで、自分の表現したい内容を伝達することができ、患者のコミュニケーションの機会が増す。

また、キーワードが不足する場合も、上記の補完技術を用いたり、複数のテキストを提示し、患者が選択することで、十分に効果を奏することができる。

人間と対話する装置に組み込むことによりコミュニケーションを助けることにも利用できる。すなわち、人間の発話文から適当にキーワードを抜き出して新たに文を作り、言い返す。文を生成したとき、典型的な情報、例えば５Ｗ１Ｈの情報などが抜けていることが分かれば、「いつ行ったのですか？」のようにその部分を尋ねる文を作るということも考えられる。

類似の構成で、音声認識して自然な文を生成し、聞き直すシステムとして提供することもできる。人間ははっきりと聞き取っているわけではなく、聞き取れなかった部分を補完して理解している。認識できた部分をもとに文を生成し、聞き直す。間違っている部分は強調して発話し直してくれることが期待できるので、何度かやり取りをする間に正しい文が生成される。

また、上記の補完技術を組み合わせて、新しいストーリーを自動的に作出するシステムを実現してもよい。例えば、「おじいさん・おばあさん・山・亀」が入力されたときに、桃太郎と浦島太郎の昔話を少なくともデータベースに備えることにより、両昔話に類似するもののそれらとは異なった新しいストーリーが作出できる。この場合、新たに補完される単語で、キーワードとして再構成されるものとしては「川・桃・竜宮城」などが考えられる。

特に、データベースに備えるストーリーが多くなればなるほど、新規性に富むストーリーが作出されるため、一読しても原文との関係が分かりにくくなる。

また、文とその文内での重要キーワードを与え、そのキーワードを含み、適切な長さの文を生成することもできるので、作文システムを実現することもできる。元の文より短ければ要約となる。文に典型的な情報を付け加えてより詳細な文を生成することも考えられる。これによって、従来の要約システムとは異なり、重要なキーワードから主体的に文を生成するため、より自然な要約が得られる。

文の苦手な者が作成した冗長な文章を修正することもできるし、語句を補い、流ちょうな文に改めることもできる。

これと関連して、文体の変換に用いることも可能である。文章からキーワードを取り出し、そのキーワードを基に文章を生成し直す。基にするデータベースに依存してそのデータベースに特有の表現に書き換えられる。例えば、ある作家の小説をデータベースにすれば、その作家風に文章を書き換えることもできるようになる。

近年急速に普及した携帯端末での文章入力支援に用いると、入力のし難い携帯端末でも、読みやすい文章を作成することができる。例えば、単語をいくつか入力すると、文候補を提示し、それから選ぶことによって、人手によって作成したのと同等の文章を作成できる。入力者は単語のみを入力するだけなので、細かく文章を作成する労がない。

また、メール作成用には、データベースに使用者の実際に作成したメールを備えることで、自己の書き方が生かされた作文が可能になる。

このように、本発明では、データベースに文体や言い回しなど、様々なテキストパターンを備えることにより、自動的に生成されるテキストがそのテキストパターンを反映するため、簡便にかつ個性的なテキストを生成することも可能である。

特に、データベースに複数の特徴的なテキストパターンを有するテキストを備えたり、多種のデータベースを備え、それらを使用者が適宜指定し、切り換えることで任意のテキストパターンを示すテキストの生成が可能である。

その他、箇条書きしたメモからキーワードとして入力し、講演用の原稿を作成したり、論文を構成することもできる。また、履歴書を入力してその人の紹介文を作成することも考えられる。

本発明は、以上の構成を備えるので、次の効果を奏する。

いくつかのキーワードを入力ステップで入力し、抽出ステップでデータベースからテキストや語句を抽出する。抽出されたテキスト又は語句を組み合わせて、入力したキーワードを用いる最適なテキストを生成することができる。

抽出されたテキストを形態素解析及び構文解析し、該テキストの係り受け構造を得れば、より自然で正確なテキスト生成を実現することも可能となる。

さらに、キーワードを含む係り受け構造を形成する過程で、係り受けモデルを用いてテキスト全体の係り受け確率を求め、該確率の最大のものを最適なテキストとして生成することで、さらに自然なテキスト生成を行える。

また、従来の構成では難しかった語順についても、語順モデルを用いて、自然な文の並びとなるテキストの生成を図ることもできる。

また、テキスト生成ステップにおいて、キーワードの全ての配列について、任意の２つのキーワード間に補完すべき単語があるか否かを学習モデルを用いて判定することもできる。学習モデルにおいて補完すべき確率の高い単語から順に補完するとき、いずれのキーワード間についても補完すべき単語がない確率が最も高くなるまで繰り返すことで、好適な補完が実現できるので、与えるキーワードが少ない場合でも、自然なテキスト生成を図ることができる。

さらに、本発明によるテキスト生成方法では、データベースに特徴的なテキストパターンを有するテキストを備えるだけで、その特徴を反映したテキストの生成ができるため、読み手にとってより違和感がなく、自然なテキスト生成方法を提供することができる。

本発明は、上記のように優れたテキスト生成方法を提供するテキスト生成

装置を創出し、自然言語処理技術の向上に寄与することが出来る。

請求の範囲

1．　文又は文章のテキストを生成するテキスト生成方法であって、
　１個以上のキーワードとなる単語を入力する入力ステップ、
　該キーワードに係るテキスト又は語句を、データベースから抽出する抽出ステップ、
　抽出されたテキスト又は語句を組み合わせて、入力したキーワードを用いる最適なテキストを生成するテキスト生成ステップ
　から構成されることを特徴とするテキスト生成方法。

2．　前記抽出ステップでテキストを抽出する構成であって、
　前記テキスト生成ステップにおいて、抽出されたテキストを形態素解析及び構文解析し、該テキストの係り受け構造を得ると共に、
　前記キーワードを含む係り受け構造を形成する
　ことを特徴とする請求の範囲第１項に記載のテキスト生成方法。

3．　前記テキスト生成ステップにおいて、
　前記キーワードを含む係り受け構造を形成する過程で、
　係り受けモデルを用いてテキスト全体の係り受け確率を求め、
　該確率の最大のものを最適なテキストとして生成する
　請求の範囲第２項に記載のテキスト生成方法。

4．　前記テキスト生成ステップにおいて、
　係り受け構造を形成する過程あるいは形成した後で、
　語順モデルを用いて自然な文の並びとなる最適なテキストを生成する
　請求の範囲第２項又は第３項に記載のテキスト生成方法。

5．　前記テキスト生成ステップにおいて、

前記キーワードの全ての配列について、任意の２つのキーワード間に補完すべき単語があるか否かを学習モデルを用いて判定し、

学習モデルにおいて補完すべき確率の高い単語から順に補完するとき、

該補完する単語をキーワードに編入し／せずに、

いずれのキーワード間についても補完すべき単語がない確率が最も高くなるまで繰り返す

単語補完処理を行う請求の範囲第１項ないし第４項に記載のテキスト生成方法。

６．　前記データベースに、特徴的なテキストパターンを有するテキストを備える構成において、

前記テキスト生成ステップが、

該特徴的なテキストパターンに準拠したテキストを生成する

請求の範囲第１項ないし第５項に記載のテキスト生成方法。

７．　文又は文章のテキストを生成するテキスト生成装置において、

１個以上のキーワードとなる単語を入力する入力手段、

複数のテキストから構成されるテキストデータベース、

該キーワードに係るテキスト又は語句を、該テキストデータベースから検索し、抽出する抽出手段、

抽出されたテキスト又は語句を組み合わせて、入力したキーワードを用いる最適なテキストを生成するテキスト生成手段

を備えたことを特徴とするテキスト生成装置。

８．　前記抽出手段でテキストを抽出する構成であって、

前記テキスト生成手段において、

抽出されたテキストを形態素解析及び構文解析し、該テキストの係り受け構造を得る解析手段と、

前記キーワードを含む係り受け構造を形成する係り受け構造形成手段と
を含む請求の範囲第７項に記載のテキスト生成装置。

９．　前記テキスト生成手段において、
前記係り受け構造形成手段が、係り受けモデルを用いてテキスト全体の係り受け確率を求め、
該確率の最大のものを最適なテキストとして生成する
請求の範囲第８項に記載のテキスト生成装置。

１０．　前記テキスト生成手段においてにおいて、
係り受け構造を形成する過程あるいは形成した後で、
語順モデルを用いて自然な文の並びとなる最適なテキストを生成する
請求の範囲第８項又は第９項に記載のテキスト生成方法。

１１．　前記テキスト生成手段において、
前記キーワードの全ての配列について、任意の２つのキーワード間に補完すべき単語があるか否かを学習モデルを用いて判定し、
学習モデルにおいて補完すべき確率の高い単語から順に補完するとき、
該補完する単語をキーワードに編入し／せずに、
いずれのキーワード間についても補完すべき単語がない確率が最も高くなるまで繰り返す
単語補完手段を含む請求の範囲第７項ないし第１０項に記載のテキスト生成装置。

１２．　前記データベースに、特徴的なテキストパターンを有するテキストを備える構成において、
前記テキスト生成手段が、
該特徴的なテキストパターンに準拠したテキストを生成する

請求の範囲第７項ないし第１１項に記載のテキスト生成装置。

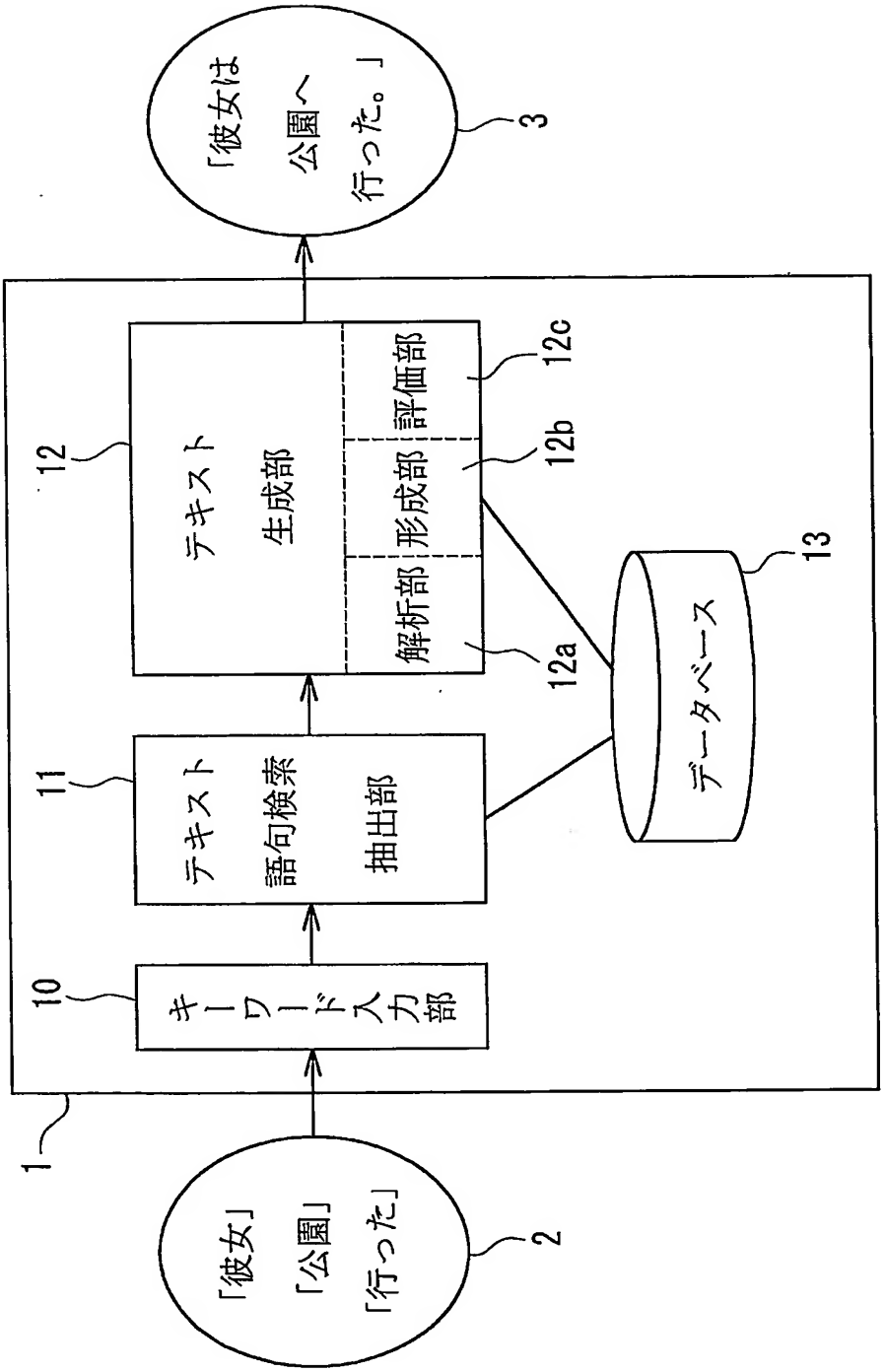
１３．　前記テキスト生成装置において、

特徴的な複数のテキストパターンを有するテキストを備えるデータベースを１つないし複数を備える一方、

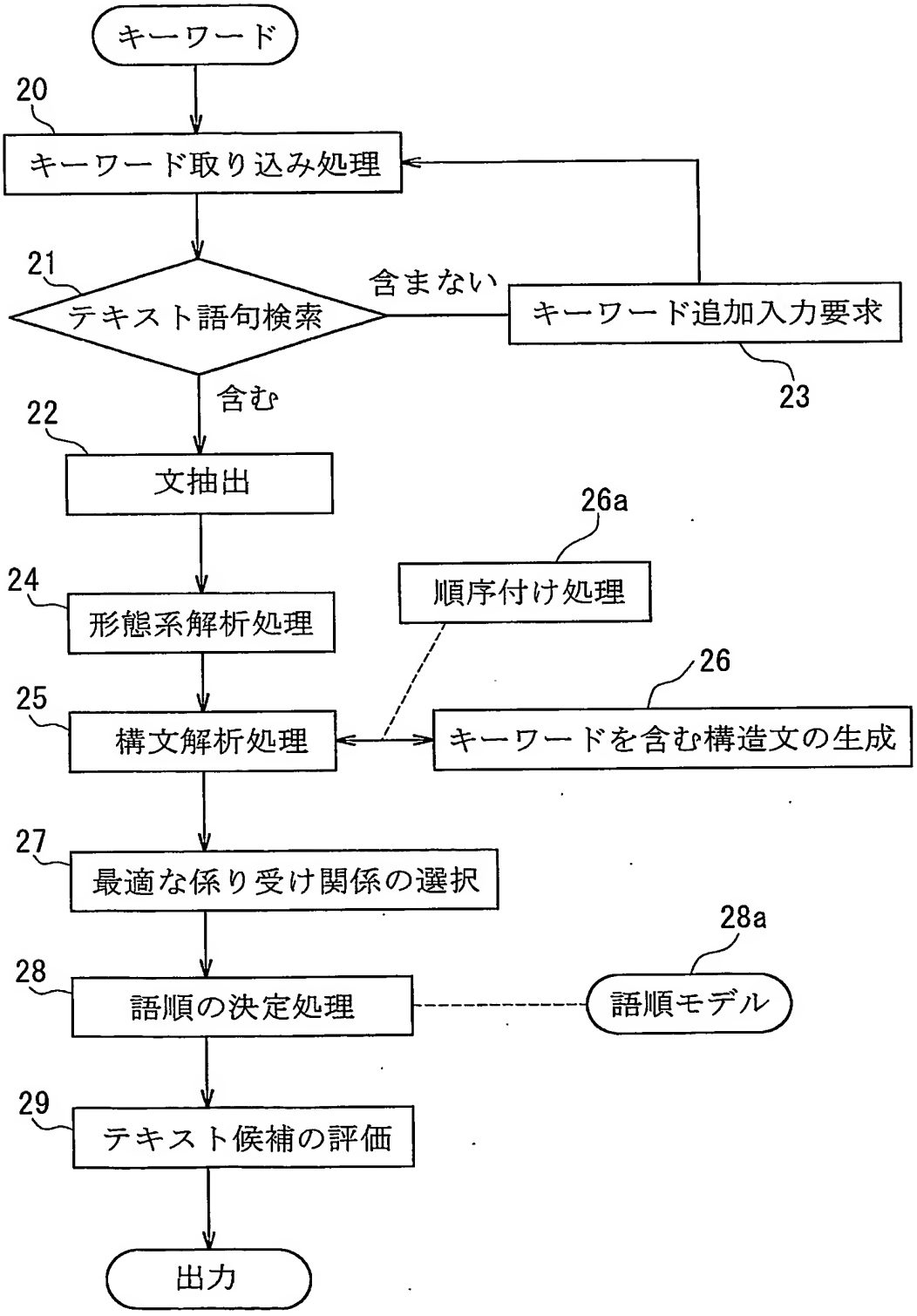
該複数のテキストパターンから所望のテキストパターンを選択するパターン選択手段を備えた

請求の範囲第１２項に記載のテキスト生成装置。

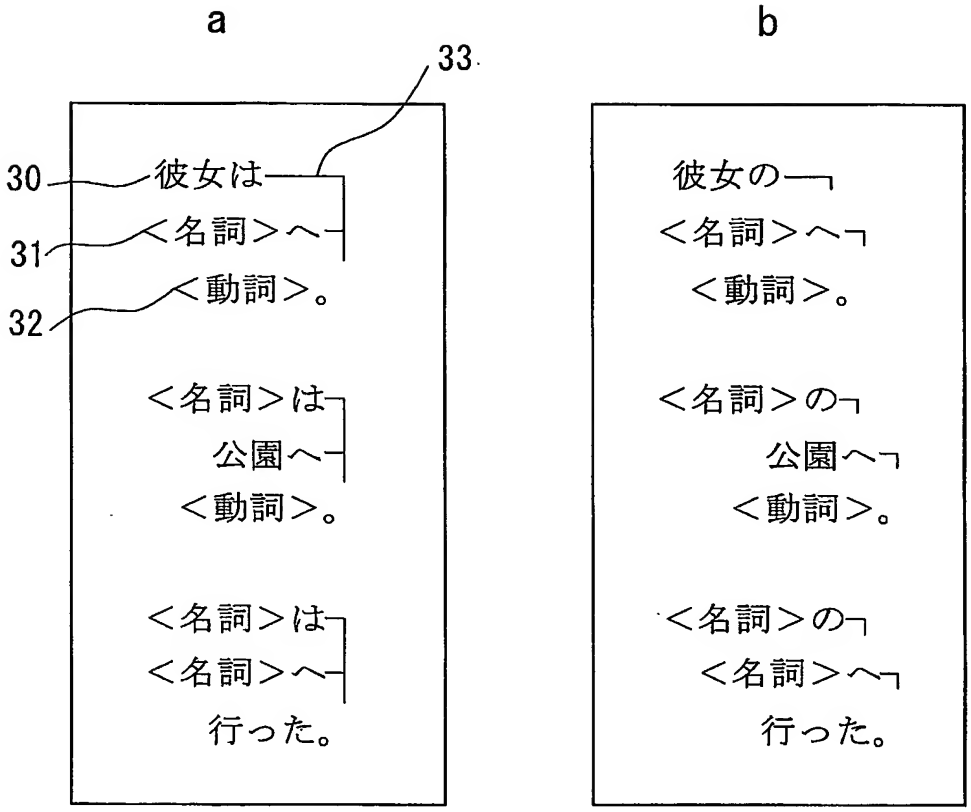
第 1 図


「彼女は
公園へ
行った。」
3
テキスト
生成部
12
解析部
形成部
評価部
12a
12b
12c
テキスト
語句検索
抽出部
11
データベース
13
キーワード入力部
10
1
「彼女」
「公園」
「行った」
2

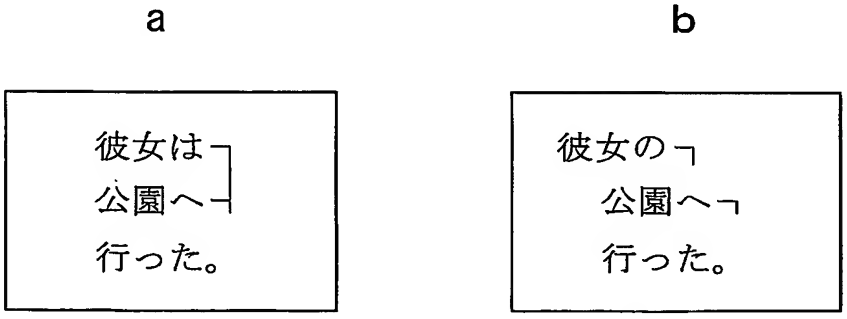
第 2 図


キーワード
20
キーワード取り込み処理
21
テキスト語句検索
含まない
キーワード追加入力要求
23
含む
22
文抽出
26a
順序付け処理
24
形態系解析処理
26
25
構文解析処理
キーワードを含む構造文の生成
27
最適な係り受け関係の選択
28a
28
語順の決定処理
語順モデル
29
テキスト候補の評価
出力

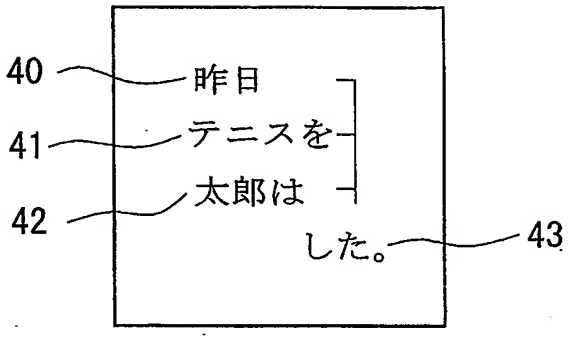
第 3 図


a
b
33
30
31
32
彼女は
<名詞>へ
<動詞>。
<名詞>は
公園へ
<動詞>。
<名詞>は
<名詞>へ
行った。
彼女の
<名詞>へ
<動詞>。
<名詞>の
公園へ
<動詞>。
<名詞>の
<名詞>へ
行った。

第 4 図


a
b
彼女は
公園へ
行った。
彼女の
公園へ
行った。

第 5 図

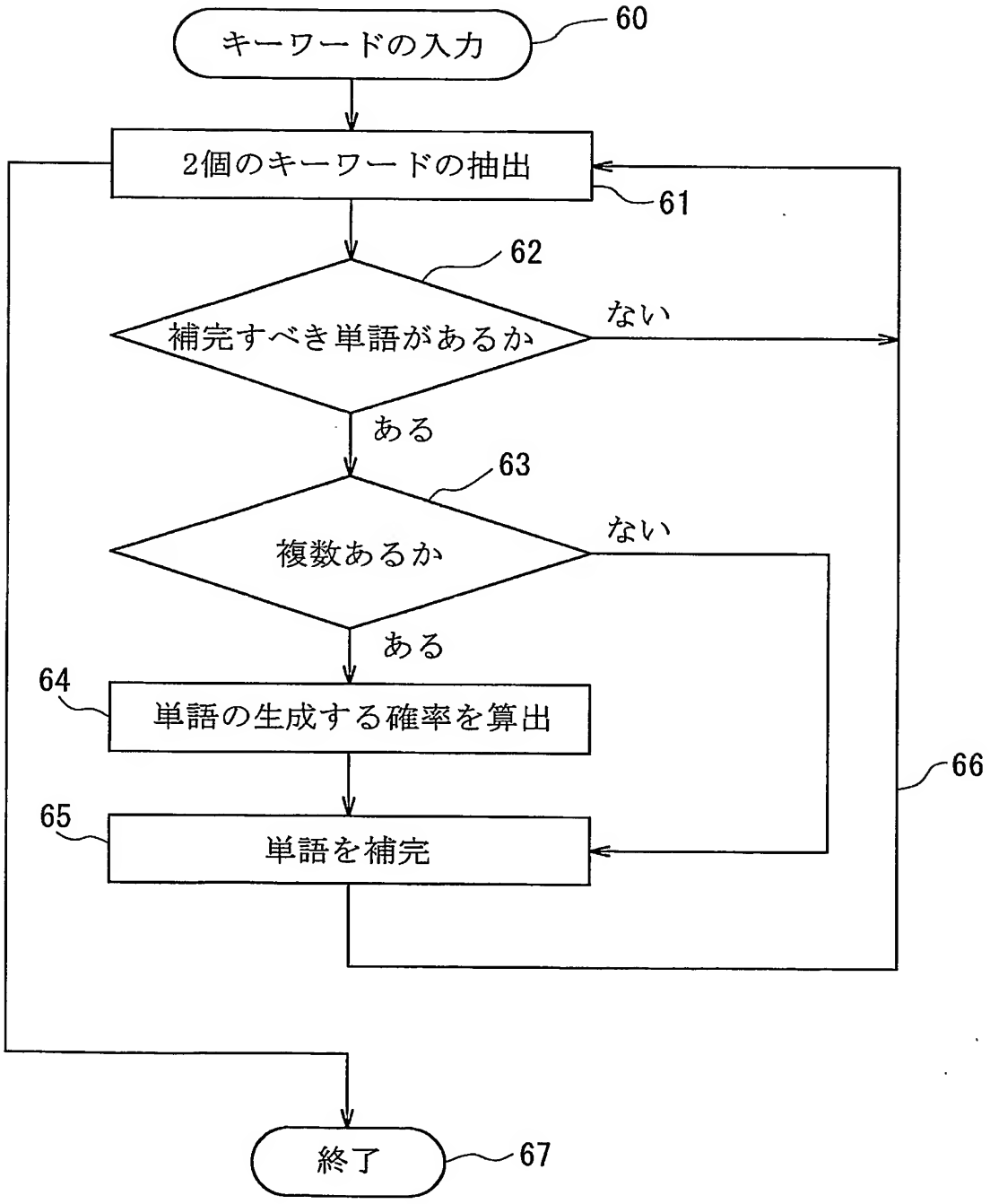


第 6 図

50

51	「昨日／太郎は／テニスを／した。」	$p^*_{昨日,太郎は} \times p^*_{昨日,テニスを} \times p^*_{太郎は,テニスを}$ $= 0.6 \times 0.8 \times 0.7 = 0.336$
52	「昨日／テニスを／太郎は／した。」	$p^*_{昨日,太郎は} \times p^*_{昨日,テニスを} \times p^*_{テニスを,太郎は}$ $= 0.6 \times 0.8 \times 0.3 = 0.144$
53	「太郎は／昨日／テニスを／した。」	$p^*_{太郎は,昨日} \times p^*_{昨日,テニスを} \times p^*_{太郎は,テニスを}$ $= 0.4 \times 0.8 \times 0.7 = 0.224$
54	「太郎は／テニスを／昨日／した。」	$p^*_{太郎は,昨日} \times p^*_{テニスを,昨日} \times p^*_{太郎は,テニスを}$ $= 0.4 \times 0.2 \times 0.7 = 0.056$
55	「テニスを／昨日／太郎は／した。」	$p^*_{昨日,太郎は} \times p^*_{テニスを,昨日} \times p^*_{テニスを,太郎は}$ $= 0.6 \times 0.2 \times 0.3 = 0.036$
56	「テニスを／太郎は／昨日／した。」	$p^*_{太郎は,昨日} \times p^*_{テニスを,昨日} \times p^*_{テニスを,太郎は}$ $= 0.4 \times 0.2 \times 0.3 = 0.024$

第 7 図


キーワードの入力
60
2個のキーワードの抽出
61
62
補完すべき単語があるか
ない
ある
63
複数あるか
ない
ある
64
単語の生成する確率を算出
65
単語を補完
66
終了
67

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP02/13185

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F17/28

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F17/28

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Toroku Jitsuyo Shinan Koho 1994-2003
Kokai Jitsuyo Shinan Koho 1971-2003 Jitsuyo Shinan Toroku Koho 1996-2003

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JICST FILE(JOIS), WPI, INSPEC(DIALOG)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	Kiyotaka UCHIMOTO, Hitoshi ISAHARA, "Saidai Entropy Model o Mochiita Nihongo Text no Ikkan Shori", The Society for Artificial Intelligence Kenkyukai Shiryo SIG-CII-2000-NOV-09, 14 November, 2000 (14.11.00)	1-4,6,7-10, 12-13 5,11
Y A	JP 05-250407 A (Hitachi, Ltd.), 28 September, 1993 (28.09.93), Par. No. [0011] & EP 585098 A & EP 560587 A & EP 629988 A & US 5473705 A & US 5699441 A & US 5887069 A	1-4,6,7-10, 12-13 5,11
A	JP 08-249331 A (Sharp Corp.), 27 September, 1996 (27.09.96), Par. No. [0009] (Family: none)	6,12-13

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 January, 2003 (08.01.03)	21 January, 2003 (21.01.03)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (July 1998)

国際調査報告

国際出願番号　PCT／JP02／13185

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁷　G06F17／28

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁷　G06F17／28

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922－1996年
日本国公開実用新案公報	1971－2003年
日本国登録実用新案公報	1994－2003年
日本国実用新案登録公報	1996－2003年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JICSTファイル（JOIS），WPI，INSPEC（DIALOG）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y A	内元清貴・井佐原均，最大エントロピーモデルを用いた日本語テキストの一貫処理，人工知能学会研究会資料SIG－CII－2000－NOV－09，2000．11．14	1- 4, 6, 7-10, 12-13 5, 11
Y A	JP 05－250407 A(株式会社日立製作所) 1993．09．28，第11段落 & EP 585098 A & EP 560587 A & EP 629988 A & US 5473705 A & US 5699441 A & US 5887069 A	1- 4, 6, 7-10, 12-13 5, 11
A	JP 08－249331 A（シャープ株式会社） 1996．09．27，第9段落（ファミリーなし）	6, 12-13

☐ C欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの

「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「O」口頭による開示、使用、展示等に言及する文献

「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」同一パテントファミリー文献

国際調査を完了した日 08．01．03	国際調査報告の発送日 21.01.03
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員） 和田　財太　5M　9459 電話番号　03－3581－1101　内線　3597

様式PCT／ISA／210（第2ページ）（1998年7月）